The Version of Record of this manuscript has been published and is available in the *International Journal of Research & Method in Education*, 2016, http://www.tandfonline.com, DOI: 10.1080/1743727X.2016.1181163

To link to this article: http://dx.doi.org/10.1080/1743727X.2016.1181163

Do we see through their eyes? Testing a bilingual questionnaire in education research using cognitive interviews

Natia Sopromadze[a]* and Pontso Moorosi[a]

[a]*Centre for Education Studies, University of Warwick, Coventry CV4 7AL, UK*

Abstract

The paper aims to demonstrate the value of cognitive interviewing (CI) as a survey pretesting method in comparative education research. Although rarely used by education researchers, CI has been successfully applied in different disciplines to evaluate and improve question performance. The method assumes that observing people's thought processes when they answer survey questions can detect response problems and point to possible solutions. To illustrate the merits of CI, we present the findings from eight cognitive interviews, which informed the development of a bilingual English/Georgian online questionnaire. The main objectives of our CI study were to a) examine cognitive validity of survey questions, b) determine semantic equivalence of the source (English) and translated (Georgian) versions of the questionnaire, and c) establish conceptual equivalence of survey measures across two cultures. We conducted two rounds of cognitive interviews, one in each language, using a combination of think-aloud and verbal probing techniques. Our analysis suggests that CI can help to identify causes of response difficulties and develop more accurate and comparable survey measures for cross-cultural education research.

Keywords: cognitive interviews; questionnaire pretesting; education research; cross-cultural surveys; validity; equivalence

*Corresponding author. Email: n.sopromadze@warwick.ac.uk, n.sopromadze@gmail.com

Background

'The real voyage of discovery consists of not in seeking new landscapes but in having new eyes.'

Marcel Proust

It can be hard to see the world through the eyes of the researched. It is a challenge to glimpse into other people's minds, follow their thoughts and understand reasoning behind their responses. When self-administered questionnaires gather data about human perceptions and attitudes, how accurate can obtained answers be? We have no tangible proof that self-report measures truly reflect research participants' perceived meanings (Schaeffer and Presser 2003; Schwarz 1999, 2007; Tourangeau, Rips, and Rasinski 2000). If there is no certainty on whether survey items are interpreted as intended, how do we assess if we elicit answers to what we really ask? Having a shared understanding of a self-report item is essential for establishing *measurement validity* – 'the extent to which an instrument measures what it is claimed to measure' (Punch 2009, 246). Validity is central to survey research and determines accuracy and meaningfulness of research results.

The quality of measurement in education survey studies has been the subject of a long debate (Gorard, Rushforth, and Taylor 2004; Gorard 2001, 2015). Education researchers have widely used self-administered questionnaires to collect descriptive and attitudinal data about social phenomena (Cohen, Manion, and Morrison 2011; Fairbrother 2014; Hartas 2010). Heavy reliance on self-report measures has been argued to limit rigour and utility of education surveys. Further methodological considerations have surfaced when transferring questionnaires from a monocultural to a cross-cultural context. Examples of specific challenges highlighted in comparative survey research in education include questionnaire construction (Thomas 2007), instrument translation and adaptation (Andrews and Diego-Mantecón 2015), issues of

equivalence (Rutkowski and Svetina 2014), and operationalisation of 'culture' (LeTendre 2002). These concerns are not limited to the field of education, but are rather inherent in comparative survey methodology in general. Regardless of the discipline, general difficulties with international surveys lie in developing valid and equivalent measurements across diverse cultures and languages (Hambleton and Zenisky 2011; Harkness et al. 2010; Smith 2003, 2004).

The issues of measurement validity and cross-cultural comparability can be addressed to a degree by carefully designing and pretesting survey items. There are various pretesting procedures such as conventional piloting, focus groups, desk appraisal, expert review, usability testing, behaviour coding and split ballot experiments (Blake 2014). While these techniques can provide valuable input into optimising the questionnaire, they offer insufficient insight into the participant's cognitive processing of individual survey items. Understanding the sources of response problems requires direct information on how questions are experienced and interpreted. An alternative pretesting method – cognitive interviewing (CI) – is argued to fill this gap.

The primary goal of CI is 'to understand the thought processes used to answer survey questions and to use this knowledge to find better ways of constructing, formulating and asking survey questions' (DeMaio and Landreth 2004, 90). The underlying assumption of this method is that observing individuals' cognitive processes reveals whether or not questions are interpreted as intended. It identifies problematic aspects in the survey design and informs the researcher which areas require modification (Beatty and Willis 2007; Collins 2014; Priede and Farrall 2011; Willis 2005; Willson and Miller 2014). Despite the promise of CI to improve the performance of self-report measures, its use in education survey research remains sparse. As cross-cultural studies in educational contexts are increasing in number (Hambleton and Zenisky 2011), advancing pretest effectiveness of comparative research instruments becomes imperative.

Engaging with more thorough methodological procedures is needed for improving question design, assessing translation quality and attaining cross-cultural comparability.

The paper aims to demonstrate how CI as a survey evaluation tool can help education researchers navigate measurement challenges. First, we review the literature on the CI techniques and reflect on the application of the method in the education field. Then we describe how the survey scales were developed and how CI fed into the process of the bilingual questionnaire development. The paper next turns to the design of our CI study. We outline specific objectives, participant recruitment and procedural steps of cognitive testing. We then analyse the findings in relation to our aims and highlight what knowledge was obtained from CI that informed the revision of the survey items. This is followed by a discussion of the CI merits and limitations offering a reflexive account of our experience with the method. Finally, the contribution of the study is presented. We conclude with suggestions for CI to be considered when designing, adapting and testing self-report questionnaires in comparative education research.

Cognitive Interviewing (CI)

CI as a survey pretesting method was developed in the 1980s as a result of interdisciplinary collaboration between cognitive psychologists and survey methodologists (Groves et al. 2009; Miller 2014; Schwarz 2007). Its theoretical model stems from cognitive theory, which breaks down the question-response process into four stages as Figure 1 below demonstrates:

[Insert Figure 1]

Building on this model, two main techniques are commonly applied in cognitive interviews: think-aloud and verbal probing (Beatty and Willis 2007; DeMaio and Landreth 2004; Priede and Farrall 2011; Willis 2005; Willis and Miller 2011). During think-alouds participants are asked to verbalise their thoughts as they interpret survey items (‘*Please tell me what you are*

thinking while you are answering the questions'). The process is participant-driven with the interviewer's role being confined to that of a facilitator making minimal intervention. On the other hand, the probing technique is interviewer-driven and involves follow-up questions eliciting information about a potentially problematic area (Collins 2003). Probing questions can be asked either concurrently, after each question during the interview, or retrospectively, after the participant completes the entire survey. Table 1 gives examples of general and specific probes applied to cognitive pretesting.

[Insert Table 1]

In contrast to survey research that draws on a large probability sample, cognitive testing typically uses a small purposive sample to gather in-depth information about each individual's understanding of survey questions (Beatty and Willis 2007; Willis 2015a). The sample size reported in CI studies of multilingual surveys ranges from 4 participants per language group (Daouk-Öyry and McDowal 2013) to over 50 (Goerman and Caspar 2010b). While it is advised to continue testing the questionnaire until flaws with question performance are no longer found, 5-15 interviews are typical and thought to be sufficient for revealing major problems (Ahmed et al. 2009; Pan 2004; Ray-Kaeser et al. 2015; Wildy and Clarke 2009; Willis 2005).

The cognitive interview method has been successfully applied in different disciplines including health research (Buers et al. 2014; Carbone, Campbell, and Honess-Morreale 2002; Garcia 2011), second language reading psychology (Ghavamnia, Ketabi, and Tavakoli 2013; Grenfell and Harris 1999; Lee-Thompson 2008) and sport science (Dietrich and Ehrlenspiel 2010). It has also been employed in developing survey measures for specific age groups, such as school-aged children (Leary, Ice, and Cottrell 2012), adolescents (Lippman et al. 2014) and older adults (Housen et al. 2008), to ensure relevance and appropriate interpretation of questions.

Recently, CI has been extended to cross-cultural contexts to establish cultural relevance of survey measures across diverse populations (Willis 2015b; Willis and Miller 2011). In addition, the method has been applied to testing translated questionnaires to judge their equivalence across different language versions (Daouk-Öyry and McDowal 2013; Farrall et al. 2012; Goerman and Caspar 2010a, 2010b; Levin et al. 2009; Park, Sha, and Pan 2014). Finally, it should be noted that CI does not serve as a substitute for conventional piloting, rather it is conducted before 'going into the field' and is an additional major step in the process of developing and testing draft questionnaires (Ornstein 2013; Willis 2015a, 5).

Application of CI in education research

Although empirical evidence in the literature suggests that CI is likely to increase the validity and reliability of self-report data, cognitive pretesting has been rarely used in education survey research. To our knowledge, seven CI studies within the field of education have been reported in peer-reviewed journals to date. We summarise key features of these sources to show the scope and nature of the existing evidence regarding the application of CI in education research (see Table 2).

[Insert Table 2]

The current paper contributes to the above literature and aims to further promote cognitive pretesting among education survey methodologists. While adding to the existing knowledge in the field, our study differs from the previous work in several ways. First, none of the reported studies tested measures on perceived emotional leadership practices in academia or involved a sample of university faculty members. We applied the CI method to examining new self-report measures in a different educational context.

Second, only two studies used CI for evaluating a cross-cultural research instrument and their procedures were not the same as ours. Wildy and Clarke (2009) adopted CI in the final phase of the International Study of Principal Preparation (ISPP) to be conducted in 13 countries. Their CI study of 5 school principals was limited to a single context – Western Australia and a single language – English. The authors state that they shared the CI findings and guidelines with their international colleagues. However, whether the researchers in other contexts also adopted this pretesting method and how these findings informed the revisions of the English version is not reported. The other study by Andrews and Diego-Mantecón (2015) explains how an instrument developed in the Flanders was adapted through CI for use in England and Spain, but it was a post hoc (sequential) adaptation of the existing Mathematics-Related Beliefs Questionnaire (MRBQ). The source (Dutch) version was not purposefully designed for comparative research. In contrast, we developed our original measures with the goal of a cross-cultural comparison. We tested both the source and target language versions and modified the English questionnaire following the input from the Georgian cognitive interviews. In this sense, our study contributes to comparative research on cognitive pretesting, which has received little attention in the field of education.

Third, we adapted the questionnaire to a non-Indo-European language. Cognitive testing has not yet been conducted in Georgian. Our story told from English and Georgian speakers' perspectives also enriches limited CI research in an education setting in non-English languages. Finally, since large-scale pretests of translated instruments are not always feasible in the social sciences, our paper illustrates a pragmatic but thorough approach to survey adaptation, which should be within the means and time of most education researchers.

Research

Questionnaire development

The survey questionnaire tested in our CI study aimed to explore emotional leadership in higher education through cross-cultural lenses. It set out to understand how academic staff perceived the emotional dimensions of departmental leadership in English and Georgian universities. Focusing on interpersonal relationships within a department team, the survey also aimed to examine academics' social identities derived from their group memberships. We intended to collect survey data at one point in time across both contexts and subsequently draw a comparison between the two academic leadership cultures. The steps we took in the process of the questionnaire development and adaptation are presented in Figure 2.

[Insert Figure 2]

Initially, we did a comprehensive review of the literature on different emotional intelligence (EI) models and cultural value systems to ensure the questionnaire captured all the essential aspects of the phenomenon. Two theoretical models were chosen to measure the key concepts: Goleman and colleagues' (2002) EI framework and Brewer and Chen's (2007) three dimensional model of individual, relational, and collective selves. When adapting the survey measures, an attempt was made to relate them to the context of the study. We tailored the developed items to work relationships and included target-specific wording (e.g. '*I am proud to be part of my department's team'; 'I consult with my colleagues before making important work-related decisions.'*).

We designed the questionnaire from the outset to be capable of cross-cultural adaptation. The term refers to the process of adapting the instrument for use in another language and culture (Beaton et al. 2000). In order to avoid potential difficulties in English-Georgian translation, simple sentences were mostly chosen avoiding phrases with metaphorical meanings. Following

Alimo-Metcalfe and Aban-Metcalfe's (2001) approach, all the statements were phrased in the same format: a) the item addressed only one dimension (e.g. teamwork); b) the item described an observable behaviour or an inferable characteristic (e. g. relationship management capability); c) the wording of the item was positive (e.g. '*[My HoD] encourages cooperation among staff members')*. After the initial question development in English, the draft questionnaire was tested by cognitive interviews to identify problematic questions before the survey was translated into Georgian. Based on the findings from the English testing round, the questionnaire was modified and after that subjected to back-translation.

Back-translation

To ensure the accuracy of translation, two bilingual translators, both native speakers of Georgian, were approached separately to do back-translation. The back-translation technique involves a) forward translation from the source language to the target language, b) blind back-translation from the target language back to the source language, and c) assessing the equivalence of both versions (Brislin 1970; Chen and Boore 2010; Smith 2004).

Although double translation is highly recommended in cross-cultural research, it may not always detect inaccuracies and lack of readability (Daouk-Öyry and McDowal 2013; Schaffer and Riordan 2003). Brislin (1970, 186) warns against the 'seaming equivalence' in a bilingual translation noting that the grammatical structure of the source language is often kept when translating it to the target language. It simplifies its back-translation and may result in a close match, but this does not necessarily mean that the two texts are semantically equivalent. Similarly, Harkness and colleagues (2004, 456) argue that in questionnaire translation 'keeping things the same is neither always possible nor always desirable'. Symmetric translation is preferred as it stays loyal to the meaning both in the source and target language and results in a

more culturally comparable translation (Sousa and Rojjanasrirat 2011). Since back-translation alone is not considered sufficient for evaluating the equivalence of bilingual items, we turned to the second round of CI in the Georgian language.

Examining language-related differences was essential, for English morpho-syntax bears no similarity to the Georgian one. Georgian (Kartuli - ქართული) being a member of the Kartvelian (South Caucasian) family of languages has its own unique alphabet and intricate grammar that largely differs from any Indo-European tongue (Hewitt 1995). Its highly agglutinative morphology allows expressing complex ideas through combining morphemes with a root word. Person and number of subjects as well as objects, tense and voice can be all combined into a single verb (Harris 1981). For example, the verb *'ვუქივარ'* (vukivar) can be translated as *'S/he has (apparently) praised me'*. Thanks to agglutination, there is a relatively free word-order in Georgian. Although the English language also has some degree of agglutination, semantic agreement of subjects, verbs and objects requires less morphological help resulting in a more fixed sentence structure (Plank 1984).

The Georgian CI round informed further revisions of the source and target versions of the questionnaire. Finally, the revised survey was field tested with a small sample of the intended population before its actual administration.

CI study design

Objectives

We adopted an iterative research design involving two rounds of cognitive interviews. The overall purpose of the CI study was to evaluate whether the bilingual English/Georgian questionnaire functioned as intended. More specifically, we had three main objectives.

First, we aimed to examine *cognitive validity* of the survey questions. Cognitive validity relates to the way people process their thoughts, emotions and experiences as they answer survey questions (Karabenick et al. 2007; Wildy and Clarke 2009). It assesses the degree of consistency between the researcher's intended meaning and the survey user's actual interpretation of a question (Muis et al. 2014). We aimed to capture the meanings of the self-report items from the participants' perspectives to examine if they meant what we assumed they did.

The second objective of CI was to judge *semantic equivalence* of the English and Georgian versions. Semantic equivalence is concerned with the performance of the questionnaire translation. It determines whether the meaning of the survey item remains the same after translating it from the source to the target language (Beck, Bernal, and Froman 2003; Daouk-Öyry and McDowal 2013; Schaffer and Riordan 2003). With the help of CI, we intended to ensure natural wording of questions as well as consistency in interpretations across languages.

Establishing *conceptual equivalence* of survey measures was our third objective when testing the bilingual questionnaire. Conceptual equivalence refers to the extent to which theoretical constructs 'elicit the same conceptual frame of reference among diverse cultural groups' (Riordan and Vandenberg 1994, 644). In other words, CI aimed to assess whether concepts were equally applicable and meaningful in each culture to make valid comparisons.

Participants

We used a direct recruitment method to identify and purposefully select suitable participants. Academic staff members were approached through personal networking in the first author's current and previous institutions, one in England and one in Georgia. Given the time constraints, we prioritised our purposive sampling criteria. Gender, age and length of service were considered as our primary variables. These characteristics were important in understanding

emotional dynamics of leader-follower interactions that the questionnaire aimed to investigate. Accordingly, we attempted to select male and female participants who were at different stages in their academic careers and showed substantial variation in terms of their age and experience (see Table 3).

[Insert Table 3]

We acknowledge that this sample size was not enough to reveal all the potentially flawed items. Nonetheless, CI sampling decisions are argued to be guided by the nature of survey questions and the aims of the study rather than a numerical goal (Willson and Miller 2014). Since the target survey questions were adapted from the existing research instruments in organisational behaviour research, they were expected to require less pretesting (Willis 2015a). Secondly, we maximized sampling efficiency by identifying a variety of people who reflected diverse experiences of the population of interest. A wide range of participants is recommended for CI as it allows examining differences in question interpretation (Beatty and Willis 2007; Collins and Gray 2014; Willis 2005). Furthermore, selecting interviewees with higher levels of education is advised as they find it easier to detect and articulate potential problems (Ackermann and Blair 2006; Collins 2014; Park, Sha, and Olmsted 2016). Considering that all our sample members were academics with higher degrees, their analytic skills were likely to facilitate problem identification. Therefore, the sample composition was deemed appropriate to uncover critical flaws with the questionnaire.

Procedure

Procedural consistency was maintained across both sample groups in terms of the administration mode and format. Following the Cross-Cultural Survey Guidelines (Survey Research Centre 2011), we pretested the draft questions in the same mode as they would be presented to the actual

survey population. Since the target questionnaire was web-based, a computerized administration mode was adopted. The research participants were provided with a laptop and a test link to the online survey. It was decided to use a mix of strategies, think-aloud and concurrent verbal probing to test 28 survey questions. The interview guide was developed in English first and then translated into Georgian. It included a set of general (participant-driven) and specific (theory-driven) pre-scripted probes to explore possible problems in the four stages of the response process (see Table 1). We also aimed to observe the participants during the survey completion to apply spontaneous (unscripted) probes, for example: *'I noticed you changed your answer from "X" to "Y". What were you thinking about?'*

One bilingual researcher, familiar with the survey topic and with experience of questionnaire design and field-based interviewing, conducted all the interviews in the same format. Before undertaking actual cognitive testing, she carried out procedural pretests of CI in both languages. The interviews were carried out in a quiet environment comfortable to the participants (e. g. university seminar rooms, participants' homes). The interviewing time varied from 60 to 90 minutes with each of the eight individuals.

First, think-aloud procedures were explained to all the research participants at the start of an interview. After practicing think-aloud with an example item, they were asked to read the questions aloud off the computer screen and verbalise their thoughts. The rationale behind reading the questions out loud rather than silently was to provide additional subtle nuances about question comprehension. The way a question was read (sometimes more than once) or a momentary pause indicated how easily a participant understood the question. The interviews were not recorded and interpretive notes were taken in the respective language while listening to the participant's narrative. The researcher entered comments under each potentially problematic

question on a pre-designed template. The notes included details about participants' task comprehension and short verbatim quotes. The interviewer's observations, such as hesitating, re-reading a question, or changing an answer, were also recorded on the same form.

Data analysis

We took a Text Summary approach to data analysis, which attempts to identify 'dominant themes, conclusions, and problems that are evidenced within a set of aggregated interviewer notes' (Willis 2015a, 60). To analyse the interview summaries systematically and compare the findings across cases, general codes were assigned to potentially flawed items. We developed a simple coding scheme from the existing error source typologies for cross-cultural CI (Fitzgerald et al. 2011; Willis and Zahnd 2007). In line with our testing objectives, we classified response difficulties into the following categories: a) cognitive, b) linguistic, c) cultural, and d) general. While these categories were rather broad, codes were supplemented with rich textual data about question functioning.

The analysis was conducted at three levels as advocated by Miller and colleagues (2011). The first analytic level (within-interview analysis) started during the interview itself when the researcher took notes. It continued immediately after the interview through the process of reviewing and summarising the written comments and assigning codes to problematic questions. The second layer (across interview analysis) examined (in)consistencies of interpretations across participants within each language group. In the last tier (across sub-group analysis), we focused on cultural and language-related differences to draw conclusions about question performance across different contexts. That is, analysis was carried out within individual interviews, across interviews and between the iterative rounds.

Both the English (source) and Georgian (translated) versions of the questionnaire were open for modifications. If it was apparent that a concept did not have an equivalent in the target language, then the source language form was revised. This process, referred to as *decentering,* implies equal importance of both language versions in the translation (Brislin 1970; Fujishiro et al. 2010; Harkness et al. 2010; Sousa and Rojjanasrirat 2011). It is meant to ensure that questions 'are not anchored in one language but fit equally well in all applicable languages' (Smith 2004, 447).

The decision to revise an item did not depend on the number of times the item was found problematic; rather it was evaluated based on the nature of the problem and logical judgement. As Willis (2005, 170) points out, *'problem frequency is not a measure of problem existence or seriousness'* [original emphasis]. Lee (2014, 230) agrees that sometimes even a single case may provide enough evidence about a potential error warranting 'proper' attention. For example, the participant's inability to map an answer on the response scale is thought to be a critical error. In this paper we share selected examples of problematic items that illustrate the key areas the study aimed to examine. We explain the nature of problem types and present possible solutions we found to the raised issues.

Findings

Cognitive validity

Regarding cognitive validity of the measures, an interesting finding emerged from English testing of the question about the emotional and social competences of Heads of Departments (HoDs). The item was originally phrased as follows: '*How important do you consider these competences for successful leadership?'* The response categories for each listed competence ranged from *'(1) not important'* to *'(5) very important'*. As the participants were reflecting on the

role of emotions in leadership, their thought processes did not show common understanding of the question intent. To get to the basis of their question comprehension, the researcher asked specific probes (e. g. *'What does 'leadership' mean to you in this context? Can you give me some examples of what you just said? Could you explain why you think that way?'*).

One interviewee assumed the question was directed at any kind of a leader rather than a HoD. Based on his experience, heading an academic department was not actually leadership but more of a managerial and administrative role. Another participant did not relate the concept of leadership to a leader as a single individual. He viewed it as a process shared among people working together as a team. His verbal report suggested that he was thinking about the emotional competences of both leaders and followers who make leadership happen together. It became apparent that it was not clear to the participants whose emotional intelligence the question targeted. To clarify ambiguity, after the English round, the original wording of the question was modified in the following way: *How important do you consider these competences for a Head of Department to be a successful leader?* When the translated version of the revised question was tested with the Georgian sample, it was not subject to competing interpretations.

However, another item that seemed to work well in the English round, caused difficulty in the Georgian one. It was designed to examine concern for group harmony and was worded as: '*I try to avoid disagreements with my colleagues.'* A Georgian participant asked if *'colleagues'* also implied a HoD. When the researcher returned the question if it would make any difference, the answer was positive. The participant would be less inclined to disagree with the department head as opposed to a staff member with an equal status. In the following administration of the question, the researcher probed into the identified problem. This time the participant's narrative indicated that healthy work relationships required exchange of different ideas and constructive

criticism. Therefore, he would not shy away from voicing his disagreement with people he worked with, including his HoD. Although this participant found the question clear, he was a more experienced academic unlike the former interviewed colleague. Since power relationships in a departmental culture as well as an individual's position may have caused inconsistency in interpretations, we decided to revise the item. Two statements were framed in place of one: '*I try to avoid disagreements with other staff member*s' and '*I try to avoid disagreements with my Head of Department.'* The revised question appeared to function well when tested further in subsequent two Georgian interviews.

Semantic equivalence

Georgian testing revealed scale-specific difficulties regarding semantic equivalence of the two language versions of the questionnaire. For example, a literal translation of a midpoint on the fully labelled Likert-type agreement scale was found to be problematic in Georgian. The option '*neither agree nor disagree*' was literally rendered as '*არც ვეთანხმები და არც არ ვეთანხმები*' (neither agree and nor *not agree*). A more comparable alternative was proposed to be '*არც ვეთანხმები და არც უარვყოფ*' (neither agree and nor *deny*). While this wording was not identical to the source scale label, it was agreed to sound more natural in the target language.

Another issue was observed regarding a *'don't know'* response. This option was included in the response scale of the items on department head's emotional and social competences (e.g. '*[My HoD] is good at managing his/her emotions in stressful situations.*'). The non-substantive response category was not highlighted as problematic in the English round, but proved otherwise in the Georgian one. A participant, who was not familiar enough with her HoD to assess the head's specific emotional competence (e.g. emotional self-control), expressed uncertainty rather

than no opinion when ticking a *'don't know'* response. It was suggested that it would be easier to respond if the scale gave an option of *'მიჭირს პასუხის გაცემა'* (difficult to answer) instead of *'არ ვიცი'* (don't know). Other participants did not raise this issue, but there were cases when they hesitated between the midpoint and *'don't know'*. When probed, it was acknowledged that *'მიჭირს პასუხის გაცემა'* (difficult to answer) option would sound better than stating that one had no opinion (don't know). Considering the verbal feedback and delayed response, we made the suggested change after the Georgian round. The English version of the questionnaire was also revised accordingly to match the meaning of the target version.

Georgian CI also discovered syntax errors in the translation. Certain items appeared to be translated literally (word-for-word) which failed to reflect connotative meanings of the original. For example, *'[My HoD] empowers staff by involving them in important decisions'* was translated as *'აძლიერებს თანამშრომლებს მათ მნიშვნელოვანი გადაწყვეტილებების მიღების პროცესში ჩართვით'*. Having retained the source language syntax, the translation followed the word arrangement of the original question. While the Georgian language does have 'free' word order, certain syntactic structures are not stylistically correct. The verb *'empowers'* translated as *'აძლიერებს'* (literally, *'makes stronger'*), does not fit naturally in the given sentence. Yet, the back-translator, who guessed the original meaning, produced a similar translation to the original English item. The Georgian participants noted that the sentence sounded awkward and suggested appropriate rewording. Since Georgian does not have an identical expression to *'empower'*, we rephrased the sentence leaving out this verb: *'თანამშრომლებს მნიშვნელოვანი გადაწყვეტილებების მიღების პროცესში რთავს'*

('[My HoD] involves staff in important decision-making.'). The phrasing of the source item was also modified reflecting the Georgian revision.

Conceptual equivalence

The concept of *ethnicity* was not interpreted within common frames of reference in the English and Georgian testing rounds. This demographic question was developed based on national census categories and each language version of the questionnaire listed relevant ethnic groups in the respective country. Although the question seemed straightforward to the English sample, it confused the Georgian cultural group members.

When the Georgian participants selected their ethnic group and moved on to the next field that asked to state their nationality, they got puzzled why they were asked the same question *twice.* '*Ethnic group*/*ethnicity'* in the Georgian language is often used interchangeably with '*nationality'* and the participants could not see a clear distinction between the two. It was suggested to remove either of the two questions as they were redundant. Based on the overall feedback, we decided to break down this category into relatively clear dimensions comprising a sense of 'shared belonging' as recommended by Burton and colleagues (2010, 1335). The revised demographic section included: country of origin, number of years living in England/Georgia, nationality, first language and religion. Although a lengthier alternative, multiple questions were expected to tap into the underlying construct better and apply it to the cultural groups being compared.

The interpretation of societal value questions also varied across the two samples. The participants were asked to agree or disagree on a five-point scale with a set of statements about the society they were currently living in. The following interview excerpt gives an example of a question item that turned out problematic for the English-speaking participants.

Original item: *In this society, most people feel proud of their cultural heritage.*

Participant: *Do you mean most people feel proud of their own cultural heritage? Or the UK cultural heritage? It's different. Because I think, to some extent everyone is proud of their own culture, the culture of their national origin. But they may not be so concerned about the culture of this country. Can you explain? How would you want me to answer?*

Researcher: *How would you respond if you were completing the survey on your own?*

Participant: *Well… the question says most people, so I would think of the majority of people, that is, White British. So yes, I would agree.*

While this participant associated *'most people'* with the predominant ethnic group in England (ONS 2012), others struggled to make a judgement: *'Hard to tell, there're so many cultural groups, hard to speak of 'most people' in a society… Most people in which community? Everybody's different… I don't really know.'*

We considered modifying the question but waited until the survey was tested with the Georgian sample to see if they had similar issues with defining their society. It has been recognized that 'Georgians, as is often true in Eastern Europe, have defined belonging to a nation in an ethnically exclusivist way' (Nodia 2005, 45). Although the ethnic composition of the country is diverse (NSOG 2014), all the four interviewed participants were unanimous in their interpretation of *'most people'* as Georgians. When we compared the feedback from different sample groups at the end of both CI rounds, we decided not to alter the question but revise its response scale. A *'not sure/difficult to answer'* option was added to the response scale of the societal value questions to allow survey users to express uncertainty if they had difficulty defining or relating to a wider society.

Discussion

The purpose of our iterative CI study was to evaluate, adapt and improve the bilingual English/Georgian questionnaire. Three specific objectives of CI were to assess cognitive validity, semantic equivalence, and conceptual equivalence of the comparative survey measures. We analysed summaries of individual interviews, compared them across interviewees per testing round and then extended the comparison across the two sample groups. On the basis of our findings, cognitive pretesting met each of our initial objectives.

First, regarding cognitive validity of survey questions, less obvious comprehension problems emerged through think-aloud and verbal probing. Despite an attempt to eliminate ambiguities while developing survey measures, CI showed that several items carried different meanings to different individuals. For example, the term 'leadership' needed clarification whether it related to a HoD's role or not. Exploring a range of interpretations led to changes in question wording to improve item clarity.

Second, in terms of semantic equivalence of different language versions, CI discovered semantic inconsistencies in the translated instrument. Some questions did not retain the meaning of the source text because of literal translation. Specifically, the middle category of the agreement scale 'neither agree nor disagree' did not translate well into Georgian. Following cognitive testing, awkward expressions and unnatural sentence structures in the target language were revised. It helped to overcome the limitations of back-translation and added methodological rigour to the process of questionnaire adaptation.

Third, with respect to conceptual equivalence, CI helped to compare internal meanings of concepts across culturally diverse samples and determined the equivalence of the survey constructs. For instance, disagreement arose over the field of *ethnicity,* which was presented as a

single uni-dimensional question with a pre-defined list of answer options. Based on the Georgian participants' feedback, we addressed the complexity of the multi-dimensional concept.

While the lessons learnt from two rounds of CI were valuable, we would like to outline the limitations of the method in the context of the study. For one, we recognize that the findings arising from a small purposive sample cannot be generalised to claim the effectiveness of CI in diagnosing question flaws. Rather, the presented results are intended to be illustrative of how this pretesting method can contribute to improving question quality in education survey research. Neither do we suggest that the conducted interviews provided comprehensive evaluation of the tested items. Willis (2015a) argues that small samples do not produce quantifiable data and increase reliance on personal judgement whether or not to revise an item. However, he also adds that certain problems are 'sample independent' and the degree of problem severity can justify the revision informed by a single participant's difficulty (Willis 2015a, 145). When interpreting the meaningfulness of individual interviews, we drew on our qualitative research skills, experience in questionnaire design and familiarity with the relevant literature on the survey topic.

We also acknowledge the criticism about artificiality of the cognitive interview process (Drennan 2003). Thinking out loud somewhat 'forces' individuals to vocalise their thoughts whereas certain cognitive processes are implicit and cannot be verbally expressed (Collins 2014; Grenfell and Harris 1999; Tourangeau, Rips, and Rasinski 2000). In addition, there is variation in individuals' articulacy, motivation to cooperate and concentration on task. In our interviews some participants were not always willing to talk through the cognitive steps occurring in their mind. In this case, the researcher would apply verbal probing to elicit relevant information (e.g. *'Can you explain why you chose this answer? Why do you say that?'*). As Willson and Miller (2014) note, sometimes interviewers have to take on a more obtrusive role to understand better

the reasoning behind survey responses. Consistent with the literature (Buers et al. 2014; Priede and Farrall 2011), we found that combining think-aloud with verbal probing maximised the effectiveness of the method.

Additionally, irrespective of the participant's degree of articulacy, there is still a risk that the researcher may either fail to detect questionnaire flaws or may identify problems which are not actually 'real' (Beatty and Willis 2007, 303). Even when a potential source of error is discovered, it is different from 'repairing' it (Willis, DeMaio, and Harris-Kojetin 1999). Then it rests on the survey designer's competence to fix it. It has been recognised that 'cognitive interviewing does not provide quantitative evidence on whether the revised version of the question proposed after cognitive testing is better than the original' (Collins 2014, 20). However, iterative rounds of cognitive testing can indicate whether new/revised items perform as expected (Watt et al. 2008; Willis 2015a). In our CI study, ideally, both language versions could have been subsequently retested to gather feedback on the modified items. We could not proceed with more rounds due to limited time. Instead, cognitive pretesting was complemented by traditional field piloting, which painted a more detailed picture of how the questionnaire functioned.

Since CI cannot be used to assess the overall time the actual survey would take participants to complete (Collins 2014), piloting gathered this information. Administering the survey online collected the timing data automatically. It allowed the evaluation of the length and flow of the whole questionnaire and its potential burden on the survey user. Combining two pretesting methods provided enough level of evidence about question performance to finalise the research instrument. We believe that the insight gained from CI could not have been provided merely by conventional pilot testing. The analysis of the verbal reports pointed to possible roots of question problems and helped to develop more accurate and comparable measures.

Conclusions

We attempted to demonstrate the value of cognitive pretesting of survey questions in cross-cultural education research. The paper reports how the results from two rounds of CI gave us direct insight into improving the bilingual questionnaire on emotional aspects of academic leadership. The comparative analysis of the interview summaries revealed difficulties with item comprehension, translation and cultural adaptation. The contribution of the study is two-fold.

First, we adopted a survey evaluation tool that has rarely been used by education researchers. Considering a rapid increase in comparative studies in education, it is vital to establish linguistic and cultural equivalence of research instruments across diverse populations. We illustrated the potential of CI to better understand the complexity of question-response process and facilitate the adaptation of the translated instrument. Our analysis suggests how applying this technique to questionnaire pretesting could improve cross-cultural survey research in education and lead to collection of more meaningful data.

In addition, the study contributes to the emerging literature on the practices of comparative CI. There is little empirical research on the effectiveness of the method in non-English languages and cultures. We applied CI to a new context as we evaluated survey comparability between the English and Georgian versions of the questionnaire. Drawing on the analysis of the participants' responses and reactions to cognitive testing, we determined that the technique performs equally well in the Georgian language. We conclude that in the quest for better survey questions, it is important to try on new lenses and see the world through the eyes of the researched. We hope this paper will encourage education survey methodologists to adopt cognitive interviews for questionnaire development and adaptation in cross-cultural contexts.

References

Ackermann, A. C., and J. Blair. 2006. "Efficient Respondent Selection for Cognitive Interviewing." Annual Meeting of the American Association of Public Opinion Research, Hollywood, FL.

Ahmed, N., J. C. Bestall, S. A. Payne, B. Noble, and S. H. Ahmedzai. 2009. "The Use of Cognitive Interviewing Methodology in the Design and Testing of a Screening Tool for Supportive and Palliative Care Needs." *Supportive Care in Cancer* 17 (6): 665-673.

Alimo-Metcalfe, B., and R. J. Alban-Metcalfe. 2001. "The Development of a New Transformational Leadership Questionnaire." *Journal of Occupational and Organizational Psychology* 74: 1-27.

Andrews, P., and J. Diego-Mantecón. 2015. "Instrument Adaptation in Cross-Cultural Studies of Students' Mathematics-Related Beliefs: Learning from Healthcare Research." *Compare: A Journal of Comparative and International Education* 45 (4): 545-567.

Beaton, D. E., C. Bombardier, F. Guillemin, and M. B. Ferraz. 2000. "Guidelines for the Process of Cross-Cultural Adaptation of Self-Report Measures." *Spine* 25 (24): 3186-3191.

Beatty, P. C., and G. B. Willis. 2007. "Research Synthesis: The Practice of Cognitive Interviewing." *Public Opinion Quarterly* 71 (2): 287-311.

Beck, Ch. T., H. Bernal, and R. D. Froman. 2003. "Methods to Document Semantic Equivalence of a Translated Scale." *Research in Nursing & Health* 26 (1): 64-73.

Blake, M. 2014. "Other Pretesting Methods." In *Cognitive Interviewing Practice*, edited by D. Collins, 28-56. London: Sage.

Brewer, M. B., and Y.-R. Chen. 2007. "Where (Who) Are Collectives in Collectivism? Toward Conceptual Clarification of Individualism and Collectivism." *Psychological Review* 114 (1): 133-151.

Brislin, R. W. 1970. "Back-Translation for Cross-Cultural Research." *Journal of Cross-Cultural Psychology* 1 (3): 185-216.

Buers, C., M. Triemstra, E. Bloemendal, N. C. Zwijnenberg, M. Hendriks, and D. M. J. Delnoij. 2014. "The Value of Cognitive Interviewing for Optimizing a Patient Experience Survey." *International Journal of Social Research Methodology* 17 (4): 325-340.

Burton, J., A. Nandi, and L. Platt. 2010. "Measuring Ethnicity: Challenges and Opportunities for Survey Research." *Ethnic and Racial Studies* 33 (8): 1332-1349.

Carbone, E. T., M. K. Campbell, and L. Honess-Morreale. 2002. "Use of Cognitive Interview Techniques in the Development of Nutrition Surveys and Interactive Nutrition Messages for Low-Income Populations." *Journal of the American Dietetic Association* 102 (5): 690-696.

Chen, H. Y., and J. R. Boore. 2010. "Translation and Back-Translation in Qualitative Nursing Research: Methodological Review." *Journal of Clinical Nursing* 19 (1-2): 234-239.

Cohen, L., L. Manion, and K. Morrison. 2011. *Research Methods in Education*. Abingdon: Routledge.

Collins, D. 2003. "Pretesting Survey Instruments: An Overview of Cognitive Methods." *Quality of Life Research* 12 (3): 229-238.

Collins, D. 2014. "Cognitive Interviewing: Origin, Purpose and Limitations." In *Cognitive Interviewing Practice*, edited by D. Collins, 3-27. London: Sage.

Collins, D., and M. Gray. 2014. "Sampling and Recruitment." In *Cognitive Interviewing Practice*, edited by D. Collins, 80-100. London: Sage.

Daouk-Öyry, L., and A. McDowal. 2013. "Using Cognitive Interviewing for the Semantic Enhancement of Multilingual Versions of Personality Questionnaires." *Journal of Personality Assessment* 95 (4): 407-416.

DeMaio, Th. J., and A. Landreth. 2004. "Do Different Cognitive Interview Techniques Produce Different Results?" In *Methods for Testing and Evaluating Survey Questionnaires*, edited by S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin, and E. Singer, 89-108. Hoboken: John Wiley & Sons.

Desimone, L. M., and K. C. Le Floch. 2004. "Are We Asking the Right Questions? Using Cognitive Interviews to Improve Surveys in Education Research." *Educational Evaluation and Policy Analysis* 26 (1): 1-22.

Dietrich, H., and F. Ehrlenspiel. 2010. "Cognitive Interviewing: A Qualitative Tool for Improving Questionnaires in Sport Science." *Measurement in Physical Education and Exercise Science* 14 (1): 51-60.

Drennan, J. 2003. "Cognitive Interviewing: Verbal Data in the Design and Pretesting of Questionnaires." *Journal of Advanced Nursing* 42 (1): 57-63.

Fairbrother, G. P. 2014. "Quantitative and Qualitative Approaches to Comparative Education." In *Comparative Education Research: Approaches and Methods*, edited by M. Bray, B. Adamson and M. Mason, 39-62. London: Springer.

Farrall, S., C. Priede, E. Ruuskanen, A. Jokinen, T. Galev, M. Arcai, and S. Maffei. 2012. "Using Cognitive Interviews to Refine Translated Survey Questions: An Example from a Cross-National Crime Survey." *International Journal of Social Research Methodology* 15 (6): 467-483.

Fitzgerald, R., S. Widdop, M. Gray, and D. Collins. 2011. "Identifying Sources of Error in Cross-National Questionnaires: Application of an Error Source Typology to Cognitive Interview Data." *Journal of Official Statistics* 27 (4): 569-599.

Fujishiro, K., F. Gong, Sh. Baron, C. J. Jacobson, Sh. DeLaney, M. Flynn, and D. E. Eggerth. 2010. "Translating Questionnaire Items for a Multi-Lingual Worker Population: The Iterative Process of Translation and Cognitive Interviews with English-, Spanish-, and Chinese-Speaking Workers." *American journal of industrial medicine* 53 (2): 194-203.

Garcia, A. A. 2011. "Cognitive Interviews to Test and Refine Questionnaires." *Public Health Nursing* 28 (5): 444-450.

Ghavamnia, M., S. Ketabi, and M. Tavakoli. 2013. "L2 Reading Strategies Used by Iranian EFL Learners: A think-Aloud Study." *Reading Psychology* 34 (4): 355-378.

Goerman, P. L., and R. A. Caspar. 2010a. "Managing the Cognitive Pretesting of Multilingual Survey Instruments: A Case Study of Pretesting of the US Census Bureau Bilingual Spanish/English Questionnaire." In *Survey Methods in Multinational, Multiregional, and Multicultural Contexts*, edited by J. A. Harkness, M. Braun, B. Edwards, T. P. Johnson, L. Lyberg, P. P. Mohler, B. E. Pennell and T. W. Smith, 75-90. Hoboken: John Wiley & Sons.

Goerman, P. L., and R. A. Caspar. 2010b. "A Preferred Approach for the Cognitive Testing of Translated Materials: Testing the Source Version as a Basis for Comparison." *International Journal of Social Research Methodology* 13 (4): 303-316.

Goleman, D., R. E. Boyatzis, and A. McKee. 2002. *Primal Leadership: Realizing the Power of Emotional Intelligence*. Boston: Harvard Business School Press.

Gorard, S. 2001. "Surveying the Field: Questionnaire Design." In *Quantitative Methods in Educational Research: The Role of Numbers Made Easy*, edited by S. Gorard, 80-108. London: Continuum.

Gorard, S. 2015. "Rethinking 'Quantitative' Methods and the Development of New researchers." *Review of Education* 3 (1): 72-96.

Gorard, S., K. Rushforth, and Ch. Taylor. 2004. "Is There a Shortage of Quantitative Work in Education Research?" *Oxford Review of Education* 30 (3): 371-395.

Greene, J. A., J. Torney-Purta, R. Azevedo, and J. Robertson. 2010. "Using Cognitive Interviewing to Explore Primary and Secondary Students' Epistemic and Ontological Cognition." In *Personal Epistemology in the Classroom: Theory, Research, and Implications for Practice*, edited by L. D. Bendixen and F. C. Feucht, 368-406. New York: Cambridge University Press.

Grenfell, M., and V. Harris. 1999. *Modern Languages and Learning Strategies: In Theory and Practice*. London: Routledge.

Groves, R. M., F. J. Fowler, M. Couper, J. M. Lepkowski, E. Singer, and R. Tourangeau. 2009. *Survey Methodology*. *Wiley Series in Survey Methodology*. Hoboken: John Wiley & Sons.

Hambleton, R. K., and A.L. Zenisky. 2011. "Translating and Adapting Tests for Cross-Cultural Assessments." In *Cross-Cultural Research Methods in Psychology*, edited by D. Matsumoto and F. J. R. Van de Vijver, 46-74. New York: Cambridge University Press.

Harkness, J. A., B. Edwards, S. E. Hansen, D. R. Miller, and A. Villar. 2010. "Designing Questionnaires for Multipopulation Research." In *Survey Methods in Multinational, Multiregional, and Multicultural Contexts*, edited by J. A Harkness, M. Braun, B. Edwards, T. P. Johnson, L. Lyberg, P. Ph. Mohler, B. Pennell and T. W. Smith, 31-57. Hoboken: John Wiley & Sons.

Harkness, J. A., B. Pennell, and A. Schoua-Glusberg. 2004. "Survey Questionnaire Translation and Assessment." In *Methods for Testing and Evaluating Survey Questionnaires*, edited by S. Presser, Rothgeb, J. M., Couper, M. P., Lessler, J. T., Martin, E., Martin, J., Singer, E., 453-473. Hoboken: John Wiley & Sons.

Harris, A. C. 1981. *Georgian Syntax: A Study in Relational Grammar*. Cambridge: Cambridge University Press.

Hartas, D. 2010. "Survey Research in Education." In *Educational Research and Inquiry: Qualitative and Quantitative Approaches*, edited by D. Hartas, 257-269. New York: Continuum International Publishing Group.

Hewitt, B. G. 1995. *Georgian: A Structural Reference Grammar*. Amsterdam and Philadelphia: John Benjamins Publishing.

Housen, P., G. R. Shannon, B. Simon, M. O. Edelen, M. P. Cadogan, L. Sohn, M. Jones, J. L. Buchanan, and D. Saliba. 2008. "What the Resident Meant to Say: Use of Cognitive Interviewing Techniques to Develop Questionnaires for Nursing Home Residents." *Gerontologist* 48 (2): 158-169.

Karabenick, S. A., M. E. Woolley, J. M. Friedel, B. V. Ammon, J. Blazevski, Ch. R. Bonney, E. Groot, M. C. Gilbert, L. Musu, and T. M. Kempler. 2007. "Cognitive Processing of Self-Report Items in Educational Research: Do They Think What We Mean?" *Educational Psychologist* 42 (3): 139-151.

Koskey, K. L., S. A. Karabenick, M. E. Woolley, Ch. R. Bonney, and B. V. Dever. 2010. "Cognitive Validity of Students' Self-reports of Classroom Mastery Goal Structure: What Students Are Thinking and Why It matters." *Contemporary Educational Psychology* 35 (4): 254-263.

Leary, J. M., Ch. Ice, and L. Cottrell. 2012. "Adaptation and Cognitive Testing of Physical Activity Measures for Use with Young, School-Aged Children and Their Parents." *Quality of Life Research* 21 (10): 1815-1828.

Lee, J. 2014. "Conducting Cognitive Interviews in Cross-National Settings." *Assessment* 21 (2): 227-240.

Lee-Thompson, L. Ch. 2008. "An Investigation of Reading Strategies Applied by American Learners of Chinese as a Foreign Language." *Foreign Language Annals* 41 (4): 702-721.

LeTendre, G. K. 2002. "Advancements in Conceptualizing and Analyzing Cultural Effects in Cross-National Studies of Educational Achievement." In *Methodological Advances in Cross-National Surveys of Educational Achievement*, A. C. Porter and A. Gamoran, 198-228. Washington, DC: National Academy Press.

Levin, K., G. B. Willis, B. H. Forsyth, A. Norberg, M. S. Kudela, D. Stark, and F. E. Thompson. 2009. "Using Cognitive Interviews to Evaluate the Spanish-language Translation of a Dietary Questionnaire." Survey Research Methods, 3 (1): 13-25.

Lippman, L. H., K. A. Moore, L. Guzman, R. Ryberg, H. McIntosh, M. F. Ramos, S. Caal, A. Carle, and M. Kuhfeld. 2014. "Cognitive Interviews: Designing Survey Questions for Adolescents." In *Flourishing Children: Defining and Testing Indicators of Positive Development*, edited by L. H. Lippman, K. A. Moore, L. Guzman, R. Ryberg, H. McIntosh, M. F. Ramos, S. Caal, A. Carle, and M. Kuhfeld. 25-43. London: Springer.

Miller, K. 2014. "Introduction." In *Cognitive Interviewing Methodology*, edited by K. Miller, V. Chepp, S. Willson and J. L. Padilla, 1-5. Hoboken: John Wiley & Sons.

Miller, K., R. Fitzgerald, J. L. Padilla, S. Willson, S. Widdop, R. Caspar, M. Dimov, et al. 2011. "Design and Analysis of Cognitive Interviews for Comparative Multinational Testing." *Field Methods* 23 (4): 379-396.

Muis, K. R., M. C. Duffy, G. Trevors, J. Ranellucci, and M. Foy. 2014. "What Were They Thinking? Using Cognitive Interviewing to Examine the Validity of Self-Reported Epistemic Beliefs." *International Education Research* 2 (1): 17-32.

Nodia, G. 2005. "Georgia: Dimensions of Insecurity." *Statehood and Security: Georgia after the Rose Revolution*, edited by B. Coppieters and R. Legvold, 39-82. Cambridge: MIT Press.

NSOG (National Statistics Office of Georgia) 2014. Statistical Yearbook of Georgia. Accessed 15 February 2016. http://geostat.ge/cms/site_images/_files/yearbook/Yearbook_2014.pdf

ONS (Office for National Statistics) 2012. Ethnicity and National Identity in England and Wales 2011. Accessed 15 February 2016. http://www.ons.gov.uk/ons/dcp171776_290558.pdf

Ornstein, M. 2013. *A Companion to Survey Research*. London: Sage.

Pan, Y. 2004. "Cognitive Interviews in Languages Other than English: Methodological and Research Issues." Proceedings of the American Statistical Association, Section on Survey Research Methods.

Park, H., M. M. Sha, and M. Olmsted. 2016. "Research Participant Selection in Non-English Language Questionnaire Pretesting: Findings from Chinese and Korean Cognitive Interviews." *Quality & Quantity*, 50 (3): 1385-1398.

Park, H., M. M. Sha, and Y. Pan. 2014. "Investigating Validity and Effectiveness of Cognitive Interviewing as a Pretesting Method for Non-English Questionnaires: Findings from

Korean Cognitive Interviews." *International Journal of Social Research Methodology* 17 (6): 643-658.

Plank, F. 1984. "Verbs and Objects in Semantic Agreement: Minor Differences Between English and German that Might Suggest a Major One." *Journal of Semantics* 3 (4): 305-360.

Priede, C., and S. Farrall. 2011. "Comparing Results from Different Styles of Cognitive Interviewing: 'Verbal Probing' vs. 'Thinking Aloud'." *International Journal of Social Research Methodology* 14 (4): 271-287.

Punch, K. 2009. *Introduction to Research Methods in Education*. Los Angeles: Sage.

Ray-Kaeser, S., T. Satink, M. Andresen, R. Martini, E. Thommen, and A. M. Bertrand. 2015. "European-French Cross-Cultural Adaptation of the Developmental Coordination Disorder Questionnaire and Pretest in French-Speaking Switzerland." *Physical & Occupational Therapy in Pediatrics* 35 (2): 132-146.

Riordan, Ch. M., and R. J. Vandenberg. 1994. "A Central Question in Cross-Cultural Research: Do Employees of Different Cultures Interpret Work-Related Measures in an Equivalent Manner?" *Journal of Management* 20 (3): 643-671.

Rutkowski, L., and D. Svetina. 2014. "Assessing the Hypothesis of Measurement Invariance in the Context of Large-Scale International Surveys." *Educational and Psychological Measurement* 74 (1): 31-57.

Schaeffer, N. Cate, and S. Presser. 2003. "The Science of Asking Questions." *Annual Review of Sociology*: 65-88.

Schaffer, B. S., and Ch. M. Riordan. 2003. "A Review of Cross-Cultural Methodologies for Organizational Research: A Best-Practices Approach." *Organizational Research Methods* 6 (2): 169-215.

Schwarz, N. 1999. "Self-Reports: How the Questions Shape the Answers." *American Psychologist* 54 (2): 93.

Schwarz, N. 2007. "Cognitive Aspects of Survey Methodology." *Applied Cognitive Psychology* 21 (2): 277-287.

Smith, T. W. 2003. "Developing Comparable Questions in Cross-National Surveys." In *Cross-Cultural Survey Methods*, edited by Harkness J. A., F. J. R. Van de Vijver and P. P. Mohler, 69-92. Hoboken: John Wiley & Sons.

Smith, T. W. 2004. "Developing and Evaluating Cross-National Survey Instruments." In *Methods for Testing and Evaluating Survey Questionnaires*, edited by S. Presser, Rothgeb, J. M., Couper, M. P., Lessler, J. T., Martin, E., Martin, J., Singer, E., 431-452. Hoboken: John Wiley & Sons.

Survey Research Center (2011). Guidelines for Best Practice in Cross-Cultural Surveys. Institute for Social Research, University of Michigan. Accessed 15 February 2016. http://ccsg.isr.umich.edu/pdf/FullGuidelines1301.pdf

Sousa, V. D., and W. Rojjanasrirat. 2011. "Translation, Adaptation and Validation of Instruments or Scales for Use in Cross-Cultural Health Care Research: A Clear and User-Friendly Guideline." *Journal of Evaluation in Clinical Practice* 17 (2): 268-274.

Thomas, A. 2007. "Self-Report Data in Cross-Cultural Research: Issues of Construct Validity in Questionnaires for Quantitative Research in Educational Leadership." *International Journal of Leadership in Education* 10 (2): 211-226.

Tourangeau, R. 1984. "Cognitive Science and Survey Methods." In *Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines*, edited by T. Jabine, M. Straf, J. Tanur and R. Tourangeau, 73-100. Washington, DC: National Academy Press.

Tourangeau, R., L. J. Rips, and K. Rasinski. 2000. *The Psychology of Survey Response*. Cambridge: Cambridge University Press.

Watt, T., Å. K. Rasmussen, M. Groenvold, J. B. Bjorner, S. H. Watt, S. J. Bonnema, L. Hegedüs, and U. Feldt-Rasmussen. 2008. "Improving a Newly Developed Patient-Reported Outcome for Thyroid Patients, Using Cognitive Interviewing." *Quality of Life Research* 17 (7): 1009-1017.

Wildy, H., and S. Clarke. 2009. "Using Cognitive Interviews to Pilot an International Survey of Principal Preparation: A Western Australian Perspective." *Educational Assessment, Evaluation and Accountability* 21 (2): 105-117.

Willis, G. B. 2005. *Cognitive Interviewing: A Tool for Improving Questionnaire Design*. Thousand Oaks: Sage.

Willis, G. B. 2015a. *Analysis of the Cognitive Interview in Questionnaire Design*. Oxford: Oxford University Press.

Willis, G. B., Th. J. DeMaio, and B. Harris-Kojetin. 1999. "Is the Bandwagon Headed to the Methodological Promised Land? Evaluating the Validity of Cognitive Interviewing

Techniques." In *Cognition and Survey Research*, edited by M. G. Sirken, D. J. Herrmann, R. Tourangeau, J. M. Tanur, N. Schwarz and S. Schechter, 133-153. New York: John Wiley & Sons.

Willis, G. B. 2015b. "The Practice of Cross-Cultural Cognitive Interviewing." *Public Opinion Quarterly* 79 (S1): 359-395.

Willis, G. B., and K. Miller. 2011. "Cross-Cultural Cognitive Interviewing: Seeking Comparability and Enhancing Understanding." *Field Methods* 23 (4): 331-341.

Willis, G. B., and E. Zahnd. 2007. "Questionnaire Design from a Cross-Cultural Perspective: An Empirical Investigation of Koreans and Non-Koreans." *Journal of Health Care for the Poor and Underserved* 18 (6): 197-217.

Willson, S., and K. Miller. 2014. "Data Collection." In *Cognitive Interviewing Methodology*, edited by K. Miller, V. Chepp, S. Willson and J. L. Padilla, 15-33. Hoboken: John Wiley & Sons.

Table 1. Examples of cognitive probes

	General	Specific
Comprehension	Can you tell me in your own words what this question is asking?	What does the term ‘empathy’ mean to you in this context?
Retrieval	How well do you recall this?	Can you remember a case when your HoD showed genuine concern for the staff members?
Judgement	How did you come up with that answer?	How accurately do you think this describes your working relationship with your HoD?
Response	How easy or difficult did you find this question to answer? Why do you say that?	Why did you choose ‘neither agree nor disagree’ and not ‘don’t know’?

Table 2. Summary of CI studies in education research

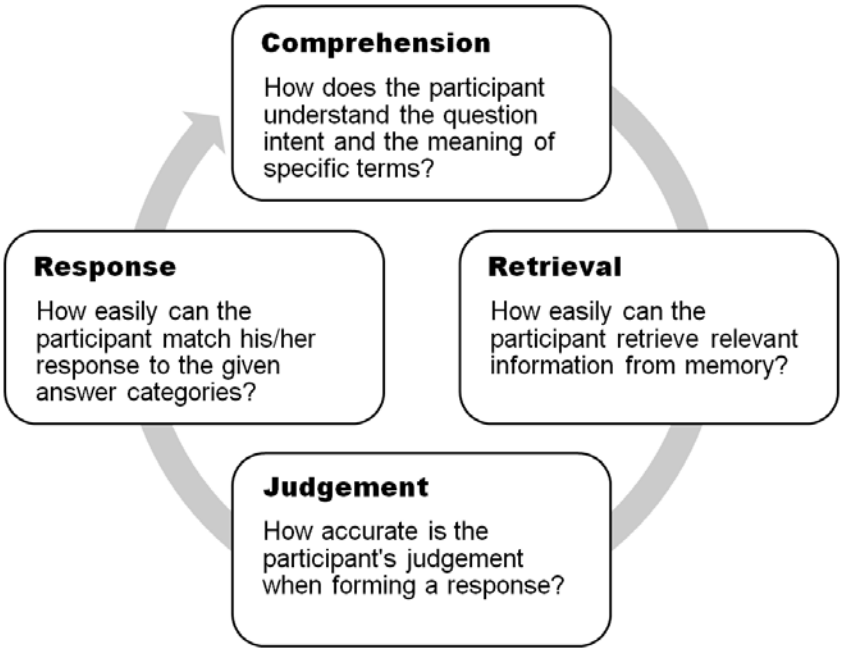
Study	Purpose	Sample	Materials tested	Procedural features
Andrews and Diego-Mantecón (2015)	Produce a cross-cultural adaptation of a questionnaire in mathematics education research	1) Secondary school students in Spain (18) 2) Secondary school students in England (18) (*n* = 36)	Mathematics-related beliefs questionnaire (MRBQ)	Think-aloud and concurrent verbal probing Duration: 30 minutes
Desimone and Le Floch (2004)	Examine validity and reliability of survey questions concerning the effects of educational reforms on classroom teaching and learning.	1) Elementary and middle school teachers (14) 2) School principals (4) (*n* = 18)	Teachers' professional development and standards-based reform surveys	Think-aloud and concurrent verbal probing Duration: 2 hours
Greene et al. (2010)	Explore students' interpretations of survey items about the nature of knowledge and knowing	1) Elementary school students (3) 2) Secondary school students (4) (*n* = 7)	Epistemic and ontological cognitions questionnaire (EOCQ)	Think-aloud, concurrent and retrospective verbal probing Duration: not stated
Karabenick et al. (2007)	Illustrate how cognitive pretesting can improve measurement validity in educational survey research.	1) Elementary school students 2) Middle school students (*n* = not stated)	Scales related to real-world instructional practices, mastery classroom goal structure, and student self-efficacy	Think-aloud and concurrent verbal probing Duration: not stated

Koskey et al. (2010)	Determine cognitive validity of students' self-report questions of classroom mastery goal structure	1) Elementary school students (19) 2) Middle school students (25) (*n* = 44)	Classroom mastery goal structure and teacher goals scales	Think-aloud and concurrent verbal probing Duration: 30 minutes
Muis et al. (2014)	Evaluate cognitive validity of a popular self-report questionnaire designed to measure students' epistemic beliefs about mathematics and psychology	1) Secondary school students (10) 2) College students (7) 3) Undergraduate students (9) 4) Graduate students (8) (*n* = 34)	Discipline-focused epistemological beliefs questionnaire (DFEBQ)	Retrospective verbal probing Duration: 1 hour
Wildy and Clarke (2009)	Pretest a cross-cultural survey with school principals to assess the appropriateness of terminology, consistency in item interpretation and question relevance to the target population	1) Novice school principals (3) 2) Experienced school principals (2) (*n* = 5)	International Study of Principal Preparation (ISSP) survey	Think-aloud and concurrent verbal probing Duration: less than 30 minutes

Table 3. CI sample composition

Participant	Gender	Age	Position	Country
01	Female	21-30	Teaching Assistant	England
02	Male	21-30	Doctoral Researcher	England
03	Male	31-40	Postdoctoral Fellow	England
04	Male	31-40	Associate Professor	England
05	Female	21-30	Doctoral Researcher	Georgia
06	Female	31-40	Lecturer	Georgia
07	Female	31-40	Assistant Professor	Georgia
08	Male	41-50	Associate Professor	Georgia

Figure 1. Four-stage response model of thought process



(Adapted from Tourangeau 1984, Willis 2005)

Figure 2. Questionnaire development process